SSOAR

Open Access Repository
www.ssoar.info

The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams

Wagner, Claudia; Singer, Philipp; Posch, Lisa; Strohmaier, Markus

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Wagner, C., Singer, P., Posch, L., & Strohmaier, M. (2013). The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds.), *The Semantic Web: Semantics and Big Data; 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013: Proceedings* (pp. 502-516). Berlin: Springer. https://doi.org/10.1007/978-3-642-38288-8_34

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-66085-7

The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams

Claudia Wagner[1], Philipp Singer[2], Lisa Posch[2], and Markus Strohmaier[2]

[1] JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
[2] Graz University of Technology, Knowledge Management Institute
Inffeldgasse 13, 8010 Graz, Austria

Abstract. Interpreting the meaning of a document represents a fundamental challenge for current semantic analysis methods. One interesting aspect mostly neglected by existing methods is that authors of a document usually assume certain background knowledge of their intended audience. Based on this knowledge, authors usually decide what to communicate and how to communicate it. Traditionally, this kind of knowledge has been elusive to semantic analysis methods. However, with the rise of social media such as Twitter, background knowledge of intended audiences (i.e., the community of potential readers) has become explicit to some extents, i.e., it can be modeled and estimated. In this paper, we (i) systematically compare different methods for estimating background knowledge of different audiences on Twitter and (ii) investigate to what extent the background knowledge of audiences is useful for interpreting the meaning of social media messages. We find that estimating the background knowledge of social media audiences may indeed be useful for interpreting the meaning of social media messages, but that its utility depends on manifested structural characteristics of message streams.

1 Introduction

In many social semantic web scenarios, understanding the meaning of social media documents is a crucial task. While existing semantic analysis methods can be used to understand and model the semantics of individual social media messages to some extent, the real time nature and the length of individual messages make it challenging to understand and model their semantics (Inches, Carman, & Crestani, 2010).

One drawback of existing methods is that they are limited to analyzing content, i.e. they do not have access to the background knowledge of potential readers. But as we know from communication theory, e.g., the Maxim of Quantity by Grice (Grice, 1975) or from Speech Act Theory (Searle, 1975), authors of messages usually make their messages as informative as required but do not provide more information than necessary. This suggests that the background knowledge of an intended audience for a given message can contribute to a semantic analysis task.

This paper sets out to study this hypothesis. We use three datasets obtained from Twitter, a popular microblogging service. Since information consumption on Twitter is mainly driven by explicitly defined social networks, we approximate the potential audience of a stream using the social network of a given author. In addition, we estimate the collective background knowledge of an audience by using the content published by the members of the audience. While the aim of this work is not to predict who will read a message, we want to approximate the collective background knowledge of a set of users who are likely to be exposed to a message and might have the background knowledge to interpret it. We do that to assess the value of background knowledge for interpreting the semantics of microblog messages. More specifically, this work addresses following research questions:

RQ1: To what extent is the background knowledge of the audience useful for guessing the meaning of social media messages? To investigate this question, we conduct a classification experiment in which we aim to classify messages into hashtag categories. As shown in (Laniado & Mika, 2010), hashtags can in part be considered as a manually constructed semantic grounding of individual microblog messages. In this work, we are going to assume that an audience which can guess the hashtag of a given message more accurately can also interprete the meaning of the message more accurately. We will use messages authored by the audience of a stream for training the classifier and we will test the performance on actual messages of a stream.

RQ2: What are the characteristics of an audience which possesses useful background knowledge for interpreting the meaning of a stream's messages and which types of streams tend to have useful audiences? To answer this question, we introduce several measures describing structural characteristics of an audience and its corresponding social stream. Then, we measure the correlation between these characteristics and the corresponding classification performance analyzed in RQ1. This shows the extent to which useful audiences can be identified based on structural characteristics.

The results of our experiments demonstrate that the background knowledge of a stream's audience is useful for the task of interpreting the meaning of microblog messages, but that the performance depends on structural characteristics of the audience and the underlying social stream. To our best knowledge, this is the first work which explores *to what extent* and *how* the background knowledge of an audience can be used to understand and model the semantics of individual microblog messages. Our work is relevant for researchers interested in learning semantic models from text and researchers interested in annotating social streams with semantics.

This paper is structured as follows: In Section 3 we give an overview about related research. Section 4 describes our experimental setup, including our methodology and a description of our datasets. Section 5 presents our experiments and empirical results. In Section 6 we discuss our results and conclude our work in Section 7.

2 Terminology

We define a *social stream* as a stream of data or content which is produced through users' activities conducted in an online social environment like Twitter where others see the manifestation of these activities. We assume that no explicitly defined rules for coordination in such environments exist. In this work we explore one special type of social streams, i.e., *hashtag streams*. A hashtag stream is a special type of a resource stream (Wagner & Strohmaier, 2010) and is defined as a tuple $S(R') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid r \in R' \vee \exists r' \in R', \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ and $R' \subseteq R$ and $Y' \subseteq Y$. In words, a hashtag stream consists of all messages containing one or several specific hashtags $r' \in R'$ and all resources (e.g., other hahstags, URLs or keywords) and users related to these messages.

In social online environments, information consumption is driven by explicitly defined social networks and therefore we can estimate the *audience* of a social stream by analyzing the incoming and outgoing links of the authors who created the stream. We call a user U_1 a *follower* of user U_2 if U_1 has established a unidirectional link with U_2 (in contrast user U_2 is a *followee* of user U_1), while we call a user U_3 a *friend* of user U_1 if U_1 has established a link with U_3 and vice versa. In this work, we assume that the union of the friends of all authors of a given hashtag constitute a hashtag stream's *audience*.

3 Related Work

Understanding and modeling the semantics of individual messages is important in order to support user in consuming social streams efficiently – e.g., via filtering social streams by users' interests or recommending tweets to users. Using topic relevance is an established approach to compute recommendations (Balabanović & Shoham, 1997) (Melville, Mooney, & Nagarajan, 2001) (Mooney & Roy, 2000).

However, the sparsity of microblog messages (i.e., the limited length of messages) makes it challenging to assess the topics of individual messages. Hence, researchers got interested in exploring the limitations of state-of-the-art text mining approaches in the context of microblogs and other short texts and develop methods for overcoming them. Two commonly used strategies for improving short text classification are: (a) improving the classifier or feature representation and (b) using background knowledge for enriching sparse textual data.

Improving the classifier or feature representation: Sriram et al. (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010) present a comparison of different text mining methods applied on individual Twitter messages. Similar to our work, they use a message classification task to evaluate the quality of the outcome of each text mining approach. Limitations of their work are that they only use 5 broad categories (news, opinions, deals, events and private message) in which they classify tweets. Further, they perform their experiments on a very small set of tweets (only 5407 tweets) which were manually assigned to the aforementioned categories. Their results show that the authorship plays a crucial role

since authors generally adhere to a specific tweeting pattern i.e., a majority of tweets from the same author tend to be within a limited set of categories. However, their authorship feature requires that tweets of the same authors occur in the trainings and test dataset.

Latent semantic models such as topic models provide a method to overcome data sparsity by introducing a latent semantic layer on top of individual documents. Hong et al. (Hong & Davison, 2010) compare the quality and effectiveness of different standard topic models in the context of social streams and examine different training strategies. To assess the quality and effectiveness of different topic models and training strategies the authors use them in two classification tasks: a user and message classification task. Their results show that the overall accuracy for classifying messages into 16 general Twitter suggest categories (e.g., Health, Food&Drinks, Books) when using topics as features is almost twice as accurate as raw TF-IDF features. Further their results suggest that the best performance can be achieved by training a topic model on aggregated messages per user. One drawback of their work is that they only use 274 users from 16 selected Twitter suggest directories[3]. These users are selected by a Twitter algorithm and it is therefore very likely that these users mainly post messages about the topic they are assigned to and that they are very popular.

In (Tang, Wang, Gao, Hu, & Liu, n.d.) the authors present an efficient approach that enriches data representation by employing machine translation to increase the number of features from different languages. Concretely the authors present a novel framework which performs multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques. The proposed approach is evaluated in terms of effectiveness on two social media datasets from Facebook and Twitter. For both Facebook and Twitter datasets, the authors construct a ground truth by selecting 30 topics from Google Trends, and retrieve the most relevant personal status or tweets via their APIs. Their results suggest that their proposed approach significantly improves the short text clustering performance.

Enriching sparse textual data with background knowledge: Based on the type of background knowledge being used, prior work can be categorized into one of the following three categories: thesaurus, web knowledge, and both of them.

Web Knowledge: Text categorization performance is improved by augmenting the bag of word representation with new features from ODP and Wikipedia as shown in (Gabrilovich & Markovitch, 2005) and (Gabrilovich & Markovitch, 2006). In (P. Wang & Domeniconi, 2008) the authors embed background knowledge derived from Wikipedia into a semantic kernel, which is then used to enrich the representation of documents. Their empirical evaluation with real data sets demonstrates that their approach successfully achieves improved classification accuracy with respect to the bag of words approach. Banerjee et al. (Banerjee, Ramanathan, & Gupta, 2007) show that clustering performance of Google news items at the feed reader end can be improved by incorporating titles of the top-

[3] http://twitter.com/invitations/suggestions

relevant Wikipedia articles as extra features. In (Phan, Nguyen, & Horiguchi, 2008) the authors present a general framework to build classifiers for short and sparse text data by using hidden topics discovered from huge text and Web collections. Their empirical results show that exploiting those hidden topics improves the accuracy significantly within two tasks: "Web search domain disambiguation" and "disease categorization for medical text".

Thesaurus or Dictionary: group words according to their similarity of meaning. Hotho et al. (Hotho, Staab, & Stumme, 2003) present an extensive study on the usage of background knowledge from WordNet for enriching documents and show that most enrichment strategies can indeed improve the document clustering accuracy. However, it is unclear if their results generalize to the social media domain since the vocabulary mismatch between WordNet and Twitter might be bigger than between WordNet and news articles.

Yoo et al. (Yoo, Hu, & Song, 2006) mapped terms in a document into MeSH concepts through the MeSH thesaurus and found that this strategy can improve the performance of text clustering. In (Shen et al., 2005) the authors use WordNet to reduce the vocabulary mismatch between the categories in the space of a search engine and the space of KDDCUP categories.

Thesaurus and Web Knowledge: For example, Hu et al. (Hu, Sun, Zhang, & Chua, 2009) cluster short texts (i.e., Google snippets) by first extracting the important phrases and expanding the feature space by adding semantically close terms or phrases from WordNet andWikipedia. Their proposed method employs a hierarchical three-level structure to tackle the data sparsity problem of original short texts and reconstruct the corresponding feature space with the integration of multiple semantic knowledge bases Wikipedia and WordNet. Empirical evaluation with Reuters and real web dataset demonstrates that their approach is able to achieve significant improvement as compared to the state-of-the-art methods.

Ontologies: include the Is-A hierarchy as well as non-taxonomic relations between entities (such as hasWonPrize).

In (Bloehdorn, Cimiano, Hotho, & Staab, 2005) the authors present an approach that uses text mining to learn the target ontology from text documents and uses then the same target ontology in order to improve the effectiveness of both supervised and unsupervised text categorization. Using Boosting as actual learning algorithm and both, term stems and concepts as features, the authors were able to achieve consistent improvements of the categorization results (1% 3% range for the Reuters-21578 corpus and in the 2.5% 7% range for the OHSUMED corpus).

In (B. B. Wang, Mckay, Abbass, & Barlow, 2002) the authors present a novel method to search for the optimal representation of a document in a domain ontology hierarchical structure to reflect concepts. Experiments have shown this is a feasible method to reduce the dimensionality of the document vector space effectively and reasonably and consequently improves the accuracy of the classifier while decreasing the computational costs. Further experiments with conceptual feature representations for supervised text categorization are presented in

(B. B. Wang, Mckay, Abbass, & Barlow, 2003) and suggest as well that concept-feature representations often outperform bag of word features.

Incorporating Background Knowledge: Hotho et al. (Hotho et al., 2003) compare several methods (add, replace, only) for incorporating background knowledge into the Bag of Words approach. The method *add* adds concepts to the word vector, while the method *replace* substitutes words with corresponding concepts. The method *only* uses only the concept vector. Hotho et al. also present different approaches for relating concepts with words. Those methods range from simple string matching to more complex word-context based disambiguation methods.

Latent semantic models such as topic models allow to incorporate background knowledge directly into the model learning step. For example, (?, ?) present approach that allows incorporating domain knowledge (in form of which words should have high or low probability in various topics) using a novel Dirichlet Forest prior in a Latent Dirichlet Allocation framework.

While (?, ?) suggest to represent background knowledge as prior probabilities of words for given topics, (?, ?) allow representing background knowledge as hierarchies of semantic concepts. In (?, ?) the authors present a probabilistic framework for combining human-defined background knowledge (represented via a hierarchy of semantic concepts) with a statistical topic model to seek the best of both worlds. Results indicate that this combination leads to systematic improvements in generalization performance.

Hashtags on Twitter: Since we use hashtags as semantic categories in which we aim to classify messages in our experiment, also research about users' hashtagging behavior is relevant for our work. In (Yang, Sun, Zhang, & Mei, 2012) the authors show that hashtags have a dual role – they are on the one hand used as topical or context marker of messages and on the other hand they are used as a symbol of community membership. The work by (Huang, Thornton, & Efthimiadis, 2010) suggests that hashtags are more commonly used to join public discussions than to organize content for future retrieval. The work of (Laniado & Mika, 2010) explores to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. Using manual annotations, they find that about half of the hashtags can be mapped to Freebase concepts with a high agreement between assessors. The authors make the assumption that hashtags are mainly used to ground tweets.

Summary: Recent research has shown promising steps towards improving short text classification by enhancing classifiers and feature representation or by using background knowledge from external sources such as Thesauri or the Web, to expand sparse textual data. However - to the best of our knowledge - using the background knowledge of intended audiences to interpret the meaning of social media messages represents a novel approach that has not been studied before. The general usefulness of such an approach is thus unknown.

4 Experimental Setup

The aim of our experiments is to explore different approaches for modeling and understanding the semantics or the main theme of microblog messages using different kinds of background knowledge. Since the audience of a microblog message are the users who are most likely to interpret (or to be able to interpret) the message, we hypothesize that the background knowledge of the audience of such messages might help to understand what a single message is about. In the following we describe our datasets and methodology.

4.1 Datasets

In this work we use three Twitter datasets each consisting of a temporal snapshot of the selected hashtag streams, the social network of stream's authors, their follower and followees and the tweets authored by the selected followers and followees (see Figure 1). We generate a diverse sample of hashtag streams as follows: In (Romero, Meeder, & Kleinberg, 2011) the authors created a classification of frequently used Twitter hashtags by category, identifying eight broad categories: celebrity, games, idioms, movies/TV, music, political, sports, and technology. We decided to reuse these categories and sample from each category 10 hashtags. We bias our random sample towards active hashtag streams by resampling hashtags for which we found less than 1,000 messages when crawling (4. March 2012). For those categories for which we could not find 10 hashtags which had more than 1,000 messages (games and celebrity) we select the most active hashtags per category (i.e., the hashtags for which we found the most messages). Since two hashtags (#bsb and #mj) appeared in the sample twice (i.e., in two different categories), we ended up having a sample of 78 different hashtags.

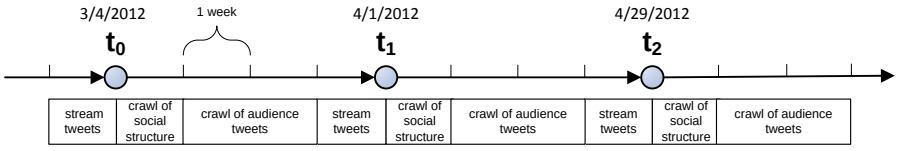


Fig. 1. Timeline of the crawling process.

Each dataset corresponds to one timeframe. The starting dates of the timeframes are March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). We crawled the most recent English tweets for each hashtag of our selection using Twitter's public search API on the first day of each timeframe and retrieved tweets that were authored within the last week. During the first week of each timeframe the user IDs of the followers and followees of streams's authors were crawled. Finally, we also crawled the most recent 3,200 tweets (or less if less were available) of

Table 1. Randomly selected hashtags per category (ordered alphabetically).

technology	idioms	sports	political	games	music	celebrity	movies
blackberry	factaboutme	f1	climate	e3	bsb	ashleytisdale	avatar
ebay	followfriday	football	gaza	games	eurovision	brazilmissesdemi	bbcqt
facebook	dontyouhate	golf	healthcare	gaming	lastfm	bsb	bones
flickr	iloveitwhen	nascar	iran	mafiawars	listeningto	michaeljackson	chuck
google	iwish	nba	mmot	mobsterworld	mj	mj	glee
iphone	nevertrust	nhl	noh8	mw2	music	niley	glennbeck
microsoft	omgfacts	redsox	obama	ps3	musicmonday	regis	movies
photoshop	oneofmyfollowers	soccer	politics	spymaster	nowplaying	teamtaylor	supernatural
socialmedia	rememberwhen	sports	teaparty	uncharted2	paramore	tilatequila	tv
twitter	wheniwaslittle	yankees	tehran	wow	snsd	weloveyoumiley	xfactor

all users who belong either to the top hundred authors or audience users of each hashtag stream. We ranked authors by the number of tweets they contributed to the stream and ranked audience users by the number of stream's authors with whom they have established a bidirectional follow relation. Figure 1 illustrates this process. Table 2 depicts the number of tweets and relations between users that we crawled during each timeframe.

Table 2. Description of the datasets.

	t_0	t_1	t_2
Stream Tweets	94,634	94,984	95,105
Audience Tweets	29,144,641	29,126,487	28,513,876
Stream Authors	53,593	54,099	53,750
Followers	56,685,755	58,822,119	66,450,378
Followees	34,025,961	34,263,129	37,674,363
Friends	21,696,134	21,914,947	24,449,705
Mean Followers per Author	1,057.71	1,087.31	1,236.29
Mean Followees per Author	634.90	633.34	700.92
Mean Friends per Author	404.83	405.09	454.88

4.2 Modeling Twitter Audiences and Background Knowledge

Audience Selection: Since the audience of a stream is potentially very large, we ranked the members of the audience according to the number of authors per stream an audience user is friend with. This allows us to determine key audience members per hashtag stream (see figure 2). We experimented with different thresholds (i.e., we used the top 10, 50 and top 100 friends) and got similar results. In the remainder of the paper, we only report the results for the best thresholds (c.f., table 3).

Background Knowledge Estimation: Beside selecting an audience of a stream, we also needed to estimate their knowledge. Hence, we compared four different methods for estimating the knowledge of a stream's audience:

– The first method (*recent*) assumes that the background knowledge of an audience can be estimated from the most recent messages authored by the audience users of a stream.

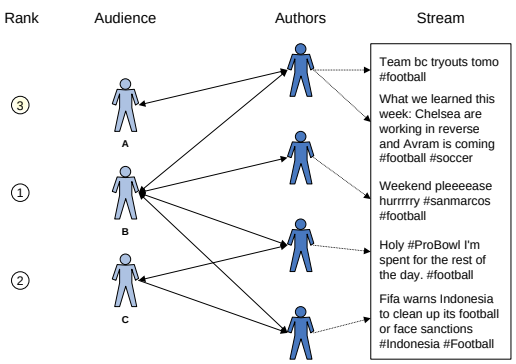


Fig. 2. To estimate the audience of a hashtag stream, we ranked the friends of the stream's authors by the number of authors they are related with. In this example, the hashtag stream #football has four authors. User B is a friend of all four authors of the stream and is therefore most likely to be exposed to the messages of the stream and to be able to interpret them. Consequently, user B receives the highest rank. User C is a friend of two authors and receives the second highest rank. The user with the lowest rank (user A) is only the friend of one author of the stream.

- The second method (*top links*) assumes that the background knowledge of the audience can be estimated from the messages authored by the audience which contain one of the top links of that audience – i.e., the links which were recently published by most audience-users of that stream. Since messages including links tend to contain only few words due to the character limitations of Twitter messages (140 characters), we test two variants of this method. In the first variant we represented the knowledge of the audience via the plain messages which contain one of the top links (*top links plain*). In the second variant (*top links enriched*) we resolved the links and enriched the messages with keywords and title information which we got from the meta-tags of the html page the links are pointing to.
- Finally, the last method (*top tags*) assumes that the knowledge of the audience can be estimated via the messages authored by the audience which contain one of the top hashtags of that audience – i.e., the hashtags which were recently used by most audience users of that stream.

4.3 Methods

In this section we present the text mining methods we used to extract content features from raw text messages. In a preprocessing step we removed all English stopwords, URLs and Twitter usernames from the content of our microblog messages. We also removed Twitter syntax such as *RT* or *via*. For stemming we used Porter Stemming. In the following part of this section we describe the text mining methods we used for producing semantic annotations of microblog messages.

Bag-of-Words Model: Vector-based methods allow us to represent each microblog message as a vector of terms. Different methods exist to weight these terms – e.g., term frequency (*TF*), inverse document frequency (*IDF*) and term frequency-inverse document frequency (*TF-IDF*). We have used different weighting approaches and have achieved the best results by using TF-IDF. Therefore, we only report results obtained from the TF-IDF weighting schema in this paper.

Topic Models: Topic models are a powerful suite of algorithms which allow discovering the hidden semantic structure in large collection of documents. The idea behind topic models is to model documents as arising from multiple topics, where each document has to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms, where few words are favored.

The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). In our experiments we used MALLET's (McCallum, 2002) LDA implementation and fitted an LDA model to our tweet corpus using individual tweets as trainings document. We chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$) and optimized them during training by using Wallach's fixed point iteration method (Wallach, 2008). We chose the number of topics T= 500 empirically by estimating the log likelihood of a model with T= 300, 500 and 700 on held out data. Given enough iterations (we used 2000) the Markov chain (which consists of topic assignments z for each token in the training corpus) has potentially converged and we can get estimates of the word distribution of topics ($\hat{\phi}$) and the topic distribution of documents ($\hat{\theta}$) by drawing samples from the chain. The estimated distributions $\hat{\phi}$ and $\hat{\theta}$ are predictive distributions and are later used to infer the topics of social stream messages.

4.4 Message Classification Task

To evaluate the quality and utility of audience's background knowledge for interpreting the meaning of microblog message, we conducted a message classification task using hashtags as classes (i.e., we had a multi-class classification problem with 78 classes). We assume that an audience which is better in guessing the hashtag of a Twitter message is better in interpreting the meaning of the message. For each hashtag stream, we created a baseline by picking the audience of another stream at random and compared the performance of the random audience with the real stream's audience. Our baseline tests how well a randomly selected audience can interpret the meaning of stream's messages. One needs to note that a simple random guesser baseline would be a weaker baseline than the one described above and would lead to a performance of 1/78.

We extracted content features (via the aforementioned methods) from messages authored by the audience of a stream before $t1$ and used them to train a classifier. That means messages of the audience of a stream were used as training samples to learn a semantic representation of messages in each hashtag class. We tested the performance of the classifier on actual messages of a stream which

were published after $t1$. In following such an approach, we ensured that our classifier does not benefit from any future information (e.g., messages published in the future or social relations which were created in the future). Out of several classification algorithms applicable for text classification such as Logistic Regression, Stochastic Gradient Descent, Multinomial Naive Bayes or Linear SVC, we could achieve the best results using a Linear SVC[4]. As evaluation metric we chose the weighted average *F1-score* which is the average of the harmonic means of precision and recall of each class weighted by the number of test samples from each class.

4.5 Structural Stream Measures

To assess the association between structural characteristics of a social stream and the usefulness of its audience (see RQ2), we introduce the following measures which describe structural aspects of those streams. We differ between static measures which only use information from one time point and dynamic measures which combine information from several time points.

Static Measures

- **Coverage Measures:** The coverage measures characterize a hashtag stream via the nature of its messages. For example the *informational coverage* measure indicates how many messages of a stream have an informational purpose - i.e., contain a link. The *conversational coverage* measures the mean number of messages of a stream that have a conversational purpose - i.e., those messages that are directed to one or several specific users. The *retweet coverage* measures the percentage of messages which are retweets. The *hashtag coverage* measures the mean number of hashtags per message in a stream.
- **Entropy Measures:** We use normalized entropy measures to capture the randomness of stream's authors and their followers, followees and friends. We rank for each hashtag stream the authors by the number of tweets they authored and the followers, followees and friends by the number of authors they are related with. A high *author entropy* indicates that the stream is created in a democratic way since all authors contribute equally much. A high *follower entropy* and *friend entropy* indicate that the followers and friends do not focus their attention towards few authors but distribute it equally across all authors. A high *followee entropy* and *friend entropy* indicate that the authors do not focus their attention on a selected part of their audience.
- **Overlap Measures:** The overlap measures describe the overlap between the authors and the followers (*Author-Follower Overlap*), followees (*Author-Followee Overlap*) or friends (*Author-Friend Overlap*) of a hashtag stream. If these overlaps are one, the stream is consumed and produced by the same users who are interconnected. A high overlap suggests that the community around the hashtag is rather closed, while a low overlap indicates that the

[4] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

community is more open and that the active and passive part of the community do not extensively overlap.

Dynamic Measures To explore how the social structure of a hashtag stream changes over time we measure the distance between the tweet-frequency distributions of stream's authors at different time points and the author-frequency distributions of stream's followers, followees or friends at different time points. We use a symmetric version of the *Kullback-Leibler (KL) divergence* which represents a natural distance measure between two probability distributions and is defined as follows: $\frac{1}{2}D_{KL}(A||B)+\frac{1}{2}D_{KL}(B||A)$. The KL divergence is *zero* if the two distributions A and B are identical and approaches infinity as they differ more and more. We measure the KL divergence for the distributions of authors, followers, followees and friends.

5 Experiments

The aim of our experiments is to explore different methods for modeling and understanding the semantics of Twitter messages using background knowledge of different kinds of audiences. Due to space restrictions we only report results obtained when training our model on the dataset t_0 and testing it on the dataset t_1. We got comparable results when training on the dataset t_1 and testing on dataset t_2.

5.1 RQ1: To what extent is the background knowledge of the audience useful for guessing the meaning of social media messages?

To answer this question we compared the performance of a classification model using messages authored by the audience of a stream (i.e., the top friends of a hashtag stream's authors) as training samples with the performance of a classification model using messages of a randomly selected audience (a baseline, i.e. the top friends of the authors of a randomly selected hashtag stream) as training samples. If the audience of a stream does not possess more knowledge about the semantics of the stream's messages than a randomly selected baseline audience, the results from both classification models should not differ significantly.

Our results show that all classifiers trained on messages authored by the audience of a hashtag stream clearly outperform a classifier trained on messages authored by a randomly selected audience. This indicates that the messages authored by the audience of a hashtag stream indeed contain important information. Our results also show that a TF-IDF based feature representation slightly outperforms a topical feature representation.

The comparison of the four different background knowledge estimation methods (see section 4.2) shows that the best results can be achieved when using the most recent messages authored by the top 10 audience users and when using messages authored by the top 100 audience users containing one of the top hashtags

of the audience (see table 3). Tweets containing one of the top links of the audience (no matter if enriched or not) are less useful than messages containing one of the top hashtags of the audience. Surprisingly, our message link enrichment strategies did not show a large boost in performance. A manual inspection of a small sample of links showed that the top links of an audience often point to multimedia sharing sites such as youtube[5], instagr.am[6] or twitpic[7]. Unfortunately, title and keywords which can be extracted from the meta information of those sites often contain information which is not descriptive.

Table 3. Average weighted F1-Scores of different classification models trained on data crawled at t_0 and tested on data crawled at t_1. We either used words weighted via TF-IDF or topics inferred via LDA as features for a message. The table shows that all audience-based classification models outperformed a random baseline. For the random baseline, we randomly swapped audiences and hashtag streams. A classifier trained on the most recent messages of the top 10 friends of a hashtag stream yields the best performance.

Classification Model	**F1 (TF-IDF)**	**F1 (LDA)**
Baseline (Random audience: top 10 friends, Messages: recent)	**0.01**	**0.01**
Audience: top 10 friends, Messages: recent	**0.25**	**0.23**
Audience: top 100 users, Messages: top links enriched	0.13	0.10
Audience: top 100 users, Messages: top links plain	0.12	0.10
Audience: top 100 users, Messages: top tags	**0.24**	**0.21**

To gain further insights into the usefulness of an audience's background knowledge, we compared the average weighted F1-Score of the eight hashtag categories from which our hashtags were initially drawn (see Table 4). Our results show that for certain categories such as sports and politics the knowledge of the audience clearly helps to learn the semantics of hashtag streams' messages, while for other streams – such as those belonging to the categories celebrities and idioms – background knowledge of the audience seems to be less useful. This suggests that only certain types of social streams are amenable to the idea of exploiting the background knowledge of stream audiences. Our intuition is that audiences of streams that are about fast-changing topics are *less useful*. We think that these audiences are only loosely associated to the topics of the stream, and therefore their background knowledge does not add much to a semantic analysis task. Analogously, we hypothesize audiences of streams that are narrow and stable are *more useful*. It seems that a community of tightly knit users is built around a topic and a common knowledge is developed over time. This seems to provide useful background knowledge to a semantic analysis task. Next, we want to understand the characteristics that distinguish audiences that are useful from audiences that are less useful.

[5] `http://www.youtube.com`
[6] `http://instagram.com/`
[7] `http://twitpic.com/`

Table 4. Average weighted F1-Score per category of the best audience-based classifier using recent messages (represented via TF-IDF weighted words or topic proportions) authored by the top ten audience users of a hashtag stream. The support represents the number of test messages for each class. We got the most accurate classification results for the category *sports* and the least accurate classification results for the category *idioms*.

		TFIDF		LDA	
category	**support**	**mean F1**	**variance F1**	**mean F1**	**variance F1**
celebrity	4384	0.17	0.08	0.15	0.16
games	6858	0.25	0.33	0.22	0.31
idioms	14562	**0.09**	0.14	**0.05**	0.05
movies	14482	0.22	0.19	0.18	0.18
music	13734	0.23	0.25	0.18	0.26
political	13200	0.36	0.22	0.33	0.21
sports	13960	**0.45**	0.19	**0.42**	0.21
technology	13878	0.22	0.20	0.22	0.2

5.2 RQ2: What are the characteristics of an audience which possesses useful knowledge for interpreting the meaning of stream's messages and which types of streams tend to have useful audiences?

To understand whether the structure of a stream has an effect on the usefulness of its audience for interpreting the meaning of its messages, we perform a correlation analysis and investigate to what extent the ability of an audience to interpret the meaning of messages correlates with structural stream properties. We use the F1-scores of the best audience based classifiers (using TFIDF and LDA) as a proxy measure for the audience's ability to interpret the meaning of stream's messages.

Figure 3a shows the strength of correlation between the F1-scores and the structural properties of streams across all categories. An inspection of the first two columns of the correlation matrix reveals interesting correlations between structural stream properties and the F1-scores of the audience-based classifiers. We further report all significant *Spearman rank correlation coefficients* ($p < 0.05$) across all categories in table 3b.

Figure 3a and table 3b show that across all categories, the measures which capture the overlap between the authors and the followers, friends and followees shows the highest positive correlation with the F1-scores. That means, the higher the overlap between authors of a stream and the followers, friends and followees of the stream, the better an audience-based classifier performs. This is not surprising since it indicates that the audience which is best in interpreting stream messages is an active audience, which also contributes to the creation of the stream itself (high author friend overlap). Further, our results suggest that the audience of a stream possesses useful knowledge for interpreting stream's messages if the authors of a stream follow each other (high author follower and author followee overlap). This means that the stream is produced and consumed by a community of users who are tightly interconnected. The only significant coverage measure is the conversational coverage measure. It indicates that the

audiences of conversational streams are better in interpreting the meaning of stream's messages. This suggests that it is not only important that a community exists around a stream, but also that the community is communicative.

All entropy measures show significant negative correlations with the F1-Scores. This shows that the more focused the author-, follower-, followee- and/or friend-distribution of a stream is (i.e., lower entropy), the higher the F1-Scores of an audience-based classification model are. The entropy measures the randomness of a random variable. For example, the author-entropy describes how random the tweeting process in a hashtag stream is – i.e., how well one can predict who will author the next message. The friend-entropy describes how random the friends of hashtag stream's authors are – i.e., how well one can predict who will be a friend of most hashtag stream's authors. Our results suggest that streams tend to have a better audience if their authors and author's followers, followees and friends are less random.

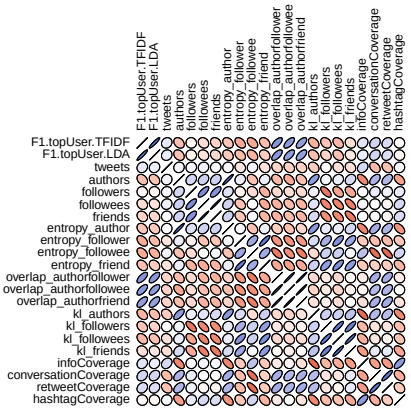
Finally, the KL divergences of the author-, follower-, and followee-distributions show a significant negative correlation with the F1-Scores. This indicates that the more stable the author, follower and followee distribution is over time, the better the audience of a stream is. If for example the followee distribution of a stream changes heavily over time, authors are shifting their social focus. If the author distribution of a stream has a high KL divergence, this indicates that the set of authors of stream are changing over time.

In summary, our results suggest that *streams which have a useful audience tend to be created and consumed by a stable and communicative community* – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

6 Discussion of Results

The results of this work show that messages authored by the audience of a hashtag stream indeed represent background knowledge that can help interpreting the meaning of streams' messages. We showed that the usefulness of an audience's background knowledge depends on the applied content selection strategies (i.e., how the potential background knowledge of an audience is estimated). However, since the audience of a hashtag stream is potentially very large, picking the right threshold for selecting the best subset of the audience is an issue. In our experiments we empirically picked the best threshold but did not conduct extensive experiments on this issue. Surprisingly, more sophisticated content selection strategies such as top links or top hashtags were only as good or even worse than the simplest strategy which used the most recent messages (up to 3,200) of each top audience user.

Our work shows that not all streams exhibit audiences which possess knowledge useful for interpreting the meaning of stream's messages (e.g., streams in certain categories like celebrities or especially idioms). Our results suggest that the utility of a stream's audience is significantly associated with structural characteristics of the stream.



(a)

feature	cor with F1 (TF-IDF)	cor with F1 (LDA)
overlap_authorfollower	0.675	0.655
overlap_authorfollowee	0.642	0.628
overlap_authorfriend	0.612	0.602
conversation_coverage	0.256	0.256
kl_followers	-0.281	–
kl_followees	-0.343	-0.302
kl_authors	-0.359	-0.307
entropy_author	-0.270	-0.400
entropy_friend	-0.307	–
entropy_follower	-0.400	-0.319
entropy_followee	-0.401	-0.368

(b)

Fig. 3. Figure 3a shows the Spearman rank correlation strength between structural stream properties and F1-Scores of two audience-based classification models averaged across all categories. The color and form of the ellipse indicate the correlation strength. Red means negative and blue means positive correlation. The rounder the ellipse the lower the correlation. The inspection of the first two columns of the correlation matrix reveals that several structural measures are correlated with the F1-Scores and table 3b shows which of those are indeed statistical significant.

Finally, our work has certain limitations. Recent research on users' hashtagging behavior (Yang et al., 2012) suggests that hashtags are not only used as topical or context marker of messages but can also be used as a symbol of community membership. In this work, we have mostly neglected the social function of hashtags. Although the content of a message may not be the only factor which influences which hashtag a user choses, we assume a "better" semantic model might be able to predict hashtags more accurately.

7 Conclusions and Future Work

This work explored whether the background knowledge of intended Twitter audiences can help in identifying the meaning of social media messages. We introduced different approaches for estimating the background knowledge of a stream's audience and presented empirical results on the usefulness of this background knowledge for interpreting the meaning of social media documents.

The main findings of our work are:

- The audience of a social stream possesses knowledge which may indeed help to interpret the meaning of stream's messages.
- The audience of a social stream is most useful for interpreting the meaning of stream's messages if the stream is created and consumed by a stable and communicative community – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

In our future work we want to explore further methods for estimating the potential background knowledge of an audience (e.g., using user lists or bio information rather than tweets). Combining latent and explicit semantic methods for estimating audience's background knowledge and exploiting it for interpreting the main theme of social media messages are promising avenues for future research.

Acknowledgments

This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner and by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz.

References

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1553374.1553378`

Balabanović, M., & Shoham, Y. (1997, March). Fab: content-based, collaborative recommendation. *Commun. ACM*, *40*(3), 66–72. Available from `http://doi.acm.org/10.1145/245108.245124`

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval* (pp. 787–788). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1277741.1277909`

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S. (2005, May). An ontology-based framework for text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*(1), 87-112.

Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 1048–1053). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Available from `http://dl.acm.org/citation.cfm?id=1642293.1642461`

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on artificial intelligence - volume 2* (pp. 1301–1306). AAAI Press. Available from `http://dl.acm.org/citation.cfm?id=1597348.1597395`

Grice, H. P. (1975). Logic and conversation. In P. Cole (Ed.), *Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the igkdd workshop on social media analytics (soma),*.

Hotho, A., Staab, S., & Stumme, G. (2003). Wordnet improves text document clustering. In *In proc. of the sigir 2003 semantic web workshop* (pp. 541–544).

Hu, X., Sun, N., Zhang, C., & Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 919–928). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1645953.1646071`

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (pp. 173–178). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1810617.1810647`

Inches, G., Carman, M., & Crestani, F. (2010). Statistics of online user-generated short documents. *Advances in Information Retrieval*, 649–652. Available from `http://dx.doi.org/10.1007/978-3-642-12275-0_68`

Laniado, D., & Mika, P. (2010). Making sense of twitter. In P. F. Patel-Schneider et al. (Eds.), *International semantic web conference (1)* (Vol. 6496, p. 470-485). Springer. Available from `http://dblp.uni-trier.de/db/conf/semweb/iswc2010-1.html#LaniadoM10`

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit.* (http://mallet.cs.umass.edu)

Melville, P., Mooney, R. J., & Nagarajan, R. (2001). Content-boosted collaborative filtering. In *In proceedings of the 2001 sigir workshop on recommender systems.*

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth acm conference on digital libraries* (pp. 195–204). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/336597.336662`

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web* (pp. 91–100). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1367497.1367510`

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 695–704). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1963405.1963503`

Searle, J. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of language* (pp. 334–369). Minneapolis: University of Minnesota Press.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., et al. (2005, December). Q2c@ust: our winning solution to query classification in kd-

dcup 2005. *SIGKDD Explor. Newsl.*, *7*(2), 100–110. Available from http://doi.acm.org/10.1145/1117454.1117467

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 841–842). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1835449.1835643

Steyvers, M., Smyth, A. P., & Chemuduganta, B. C. (2011). Combining Background Knowledge and Learned Topics. *Topics in Cognitive Science*, *3*(18–47). Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.7316

Tang, J., Wang, X., Gao, H., Hu, X., & Liu, H. (n.d.). Enriching short texts representation in microblog for clustering. *Frontiers of Computer Science*.

Wagner, C., & Strohmaier, M. (2010). The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Semantic search workshop at www2010.* Available from http://www.student.tugraz.at/claudia.wagner/publications/wagner_semsearch2010.pdf

Wallach, H. M. (2008). *Structured topic models for language.* Unpublished doctoral dissertation, University of Cambridge.

Wang, B. B., Mckay, R. I. (bob, Abbass, H. A., & Barlow, M. (2002). Learning text classifier using the domain concept hierarchy. In *In proceedings of international conference on communications, circuits and systems 2002* (pp. 1230–1234). Press.

Wang, B. B., Mckay, R. I. B., Abbass, H. A., & Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. In *Proceedings of the 26th australasian computer science conference - volume 16* (pp. 69–78). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Available from http://dl.acm.org/citation.cfm?id=783106.783115

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 713–721). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1401890.1401976

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on world wide web* (pp. 261–270). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2187836.2187872

Yoo, I., Hu, X., & Song, I.-Y. (2006). Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 791–796). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1150402.1150505